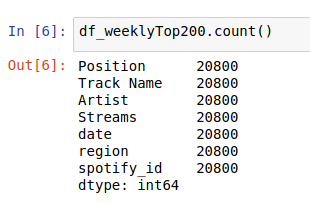
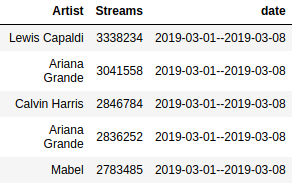
# Cleanup Steps for the Spotify datasets

## Spotify weekly Top200

* 1. Checking missing values: there is no column with null/missing values



* 1. Check valid date values: the date column is of string type populated by weekly date range (Fig 1.2a). The date column is split into start and end dates as illustrated in fig 1.2b below.



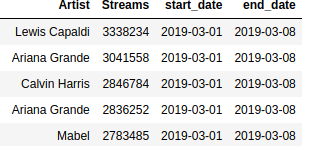
 **BEFORE AFTER**

Figure 1.2a: date column with weekly date range as string

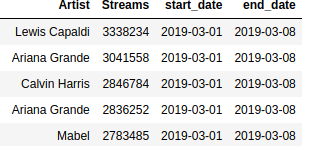


Figure 1.2b: start date and end date as separate columns

## Spotify Track Details

* 1. album release date (some of them contain just year, no day and month) (default such to 1st Jan)
  2. some artist names come in double quote ("Earth, Wind & Fire")
  3. also watch out for some artist names with comma (Earth, Wind & Fire).

**NB**

Although 2.2 & 2.3 shouldn’t be a problem to pandas, it may cause problems when importing into MS Access. The observation in 2.1 was corrected by converting the string column into date. As a result, default values such as 1990 were converted to 1990-01-01 as illustrated in figures 2.1a and 2.1b

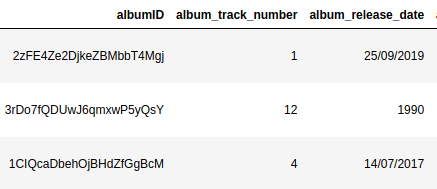
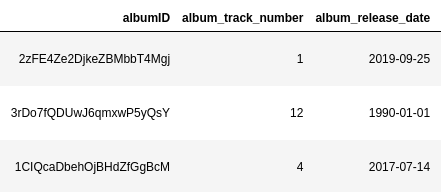
 

Figure 2.1a: album release date showing only the year

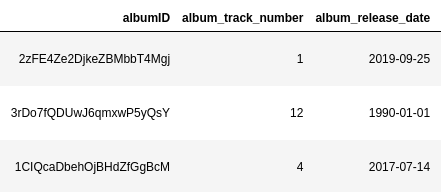


Figure 2.1b: album release date formatted

* 1. Artists without music-brainz identifier (mbid)

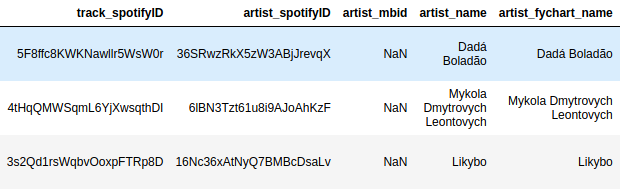


Figure 2.4: Artists without mbid